MODELING ORDINAL SOPHISTICATION OF PROBLEM-SOLVING STRATEGIES FOR LENGTH-MEASUREMENT: **THE HURDLE APPROACH**

Carson Keeter

University of Wyoming – Department of Statistics

Supervised by:

Pavel Chernyavskiy, PhD; *University of Virginia – School of Medicine*

Traci Shizu Kutaka, PhD; *University of Denver – Marsico Institute*

MOTIVATION

- What **metrics** should we use to make inferences about student learning?
- What is the **process** we use to generate these metrics?
- What **evidence** do we have that enable us to trust these metrics?

- Length-measurement is understudied even though it is a critical bridge between geometry and number concepts (Clements, 2021; Sarama, 2009).

RESEARCH QUESTIONS

1. Do children in the Learning Trajectories group use more sophisticated strategies relative to their peers in other learning approaches?

2. How do we model strategy preference, given a large portion of strategies do not fall onto the existing research-based sophistication scale?
   - Detectable Strategies *vs* Non-Codable and Non-Detectable strategies
   - What can *item*, *student*, and *classroom* random effects tell us about strategy preference?

SAMPLING METHODS AND DEMOGRAPHICS

- **Sampling** (Randomized Control Trial)
  - n = 186 kindergarten students
    - 16 classes across 6 schools
    - 149 in 4 public schools, 37 in 2 private schools
    - 104 girls, 82 boys
- **School District**
  - 53.8% Latinx, 24.7% White, 13.2% African American, 3.2% Asian, 0.7% American Indian, 0.4% Native Hawaiian or Other Pacific Islander, and 4.1% respondents who identified as having Two or More Races.
  - 65% of students qualify for free-/reduced-lunch
  - 36.3% are English-Language Learners

EXPERIMENTAL CONDITIONS

- **Three Conditions:**
  - Learning Trajectory (LT): 70 students
  - Reverse-order (REV): 59 students
  - Business-as-Usual (BAU): 57 students

- LT and REV students received one-on-one instruction using the same activities from the length LT, while the REV condition reversed the order of activities presented

- BAU did not receive one-on-one instruction

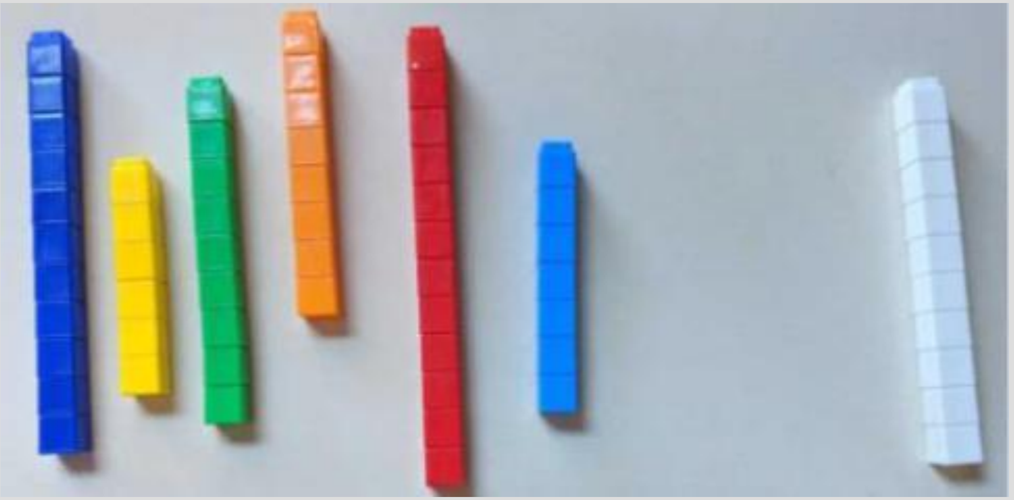
https://www.learningtrajectories.org/math/learning-trajectories

DATA COLLECTION AND CODING:
PRE- AND POST-ASSESSMENT

- Composed of 28 items developed to assess length measurement learning (*Research-Based Early Mathematics Assessment*, Clements, 2008/2020)

- Each item was scored for correctness and *sophistication for the strategy observed*

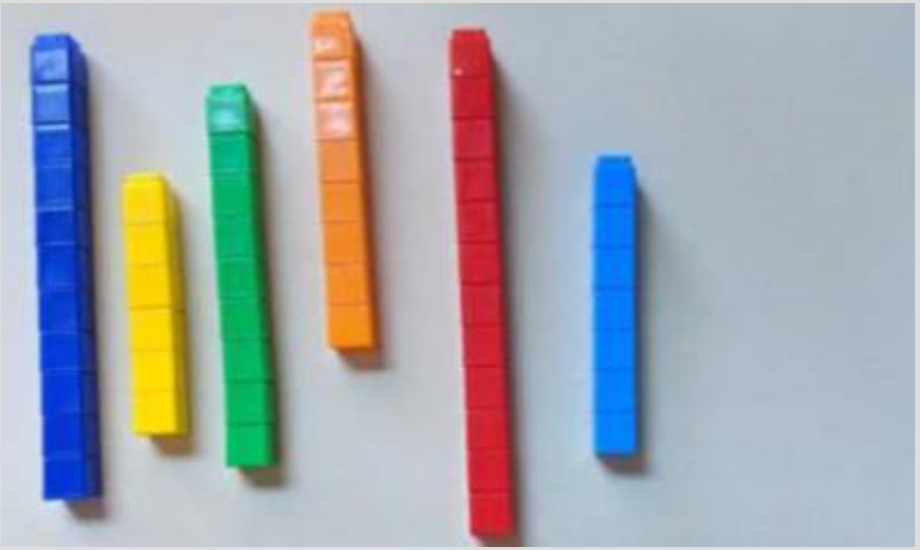
- *Sophistication* was made up of up to 10 research-based codes then collapsed into 4-point ordinal scores.

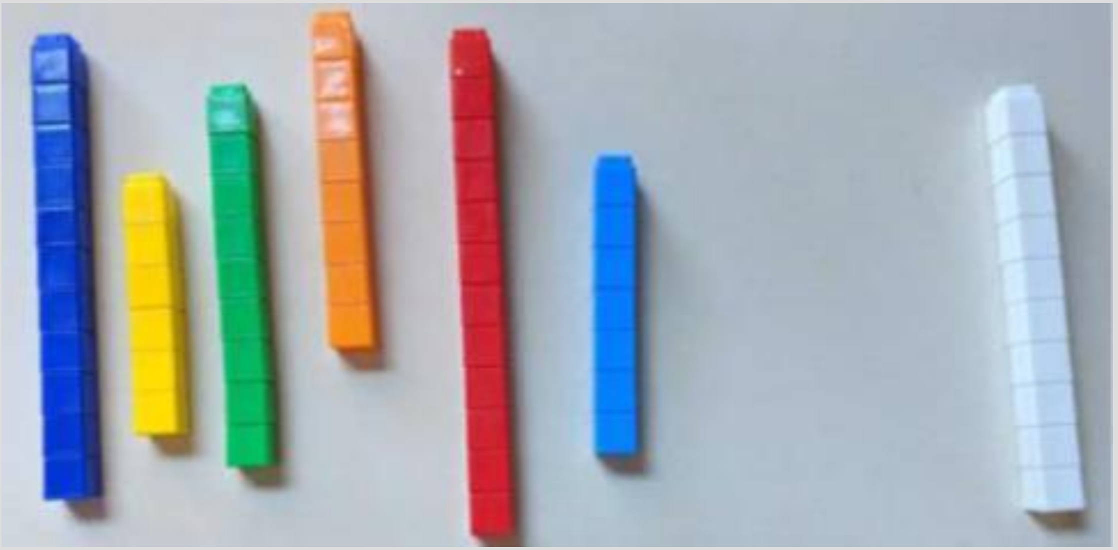


DATA COLLECTION AND CODING:
PRE- AND POST-ASSESSMENT

Example:

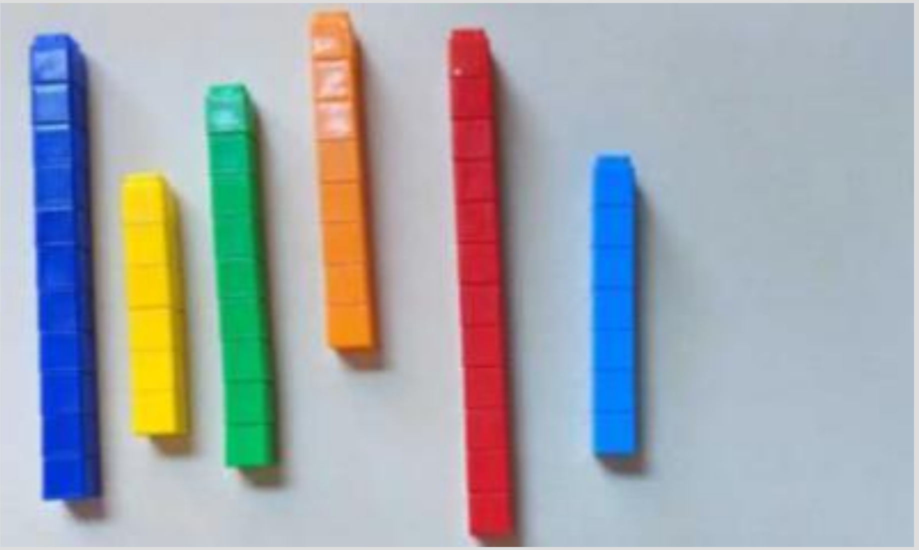
"Can you put these in order from the shortest to longest?" When the child is done, give him/her the tower made of ten white cubes, and say, **"I forgot this one! Can you place it so they are all in order?"**

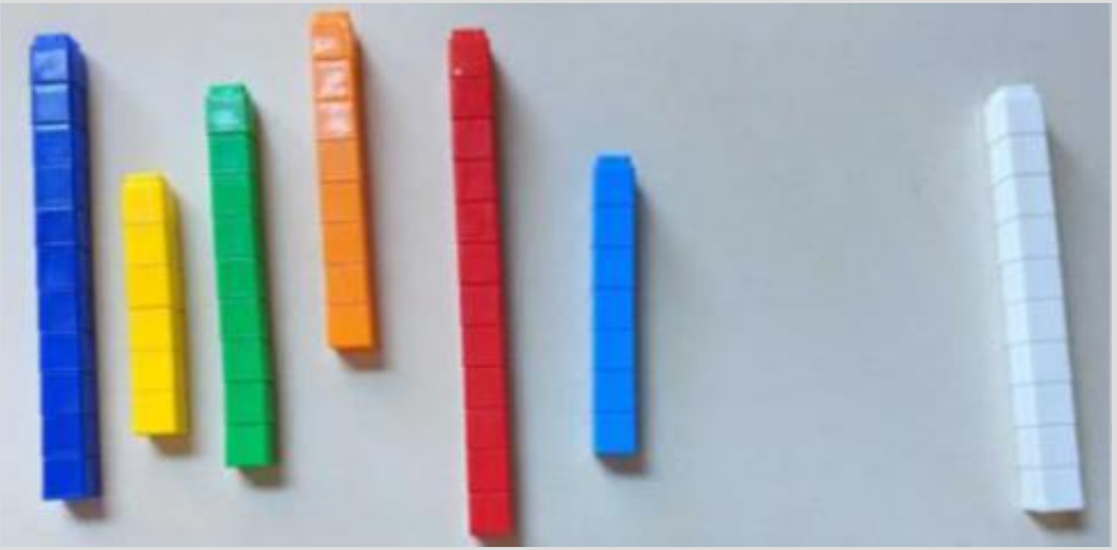




Item H from *LT Studies: Measurement* Unit 6 Length – Fall 2019

DATA COLLECTION AND CODING:
PRE- AND POST-ASSESSMENT





- Research Based Sophistication Codes:

1 = Plays without attempting item

10 = Uses trial and error repeatedly to order

2 = Compares only 2 towers at a time

3 = Separates into categorizes of length without complete ordering within them

4 = Attempts to order without alignment

5 = Attempts to order with alignment

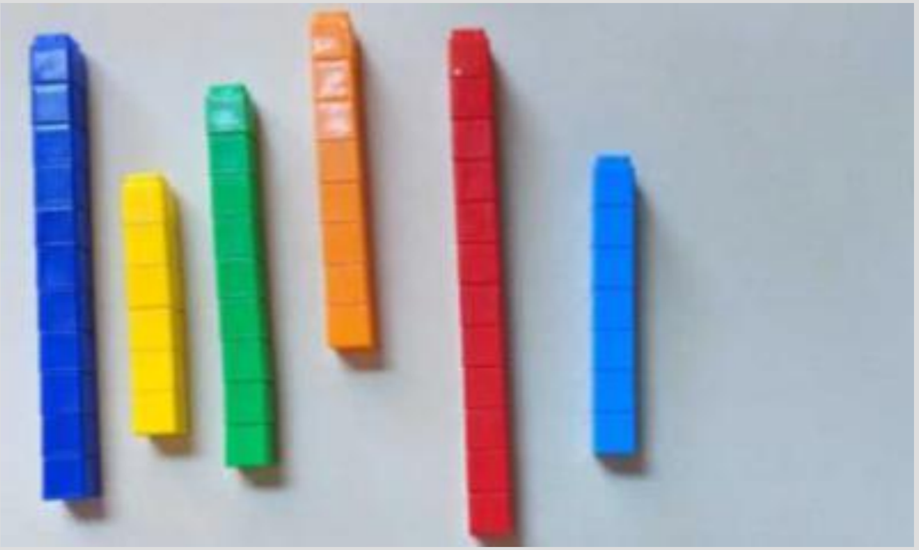
6 = Systematically searches for the next long/shorter tower

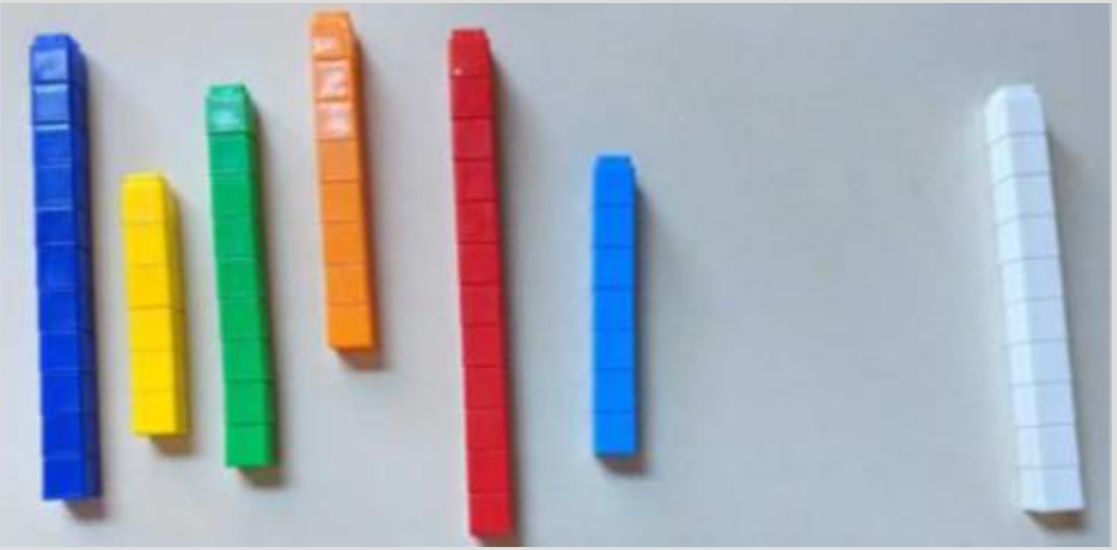
7 = Other

8 = Strategy could not be observed

9 = NA

DATA COLLECTION AND CODING:
PRE- AND POST-ASSESSMENT





- 4-Point Ordinal Sophistication Coding (collapsed):

L0: 1 = Plays without attempting item

L1: 2 = Compares only 2 towers at a time; 3 = Separates into categorizes of length without complete ordering within them; 4 = Attempts to order without alignment; 10 = Uses trial and error repeatedly to order

L2: 5 = Attempts to order with alignment

L3: 6 = Systematically searches for the next long/shorter tower

H: 7 = Other; 8 = Strategy could not be observed; 9 = NA

CODING PROCESS

- Most Detectable Strategies were coded 0, 1, 2, or 3 depending on how sophisticated the utilized strategy was
  - 0 = least sophisticated → 3 = most sophisticated
- Non-Codable and Non-Detectable Strategies were coded as *H* (to be processed in the hurdle model)
  - Cannot be placed on same sophistication scale as Detectable Strategies

L0: **1** = Plays without attempting item

L1: **2** = Compares only 2 towers at a time; **3** = Separates into categorizes of length without complete ordering within them; **4** = Attempts to order without alignment; **10** = Uses trial and error repeatedly to order

L2: **5** = Attempts to order with alignment

L3: **6** = Systematically searches for the next long/shorter tower

H: **7** = Other; **8** = Strategy could not be observed; **9** = NA

THE HURDLE MODELING APPROACH

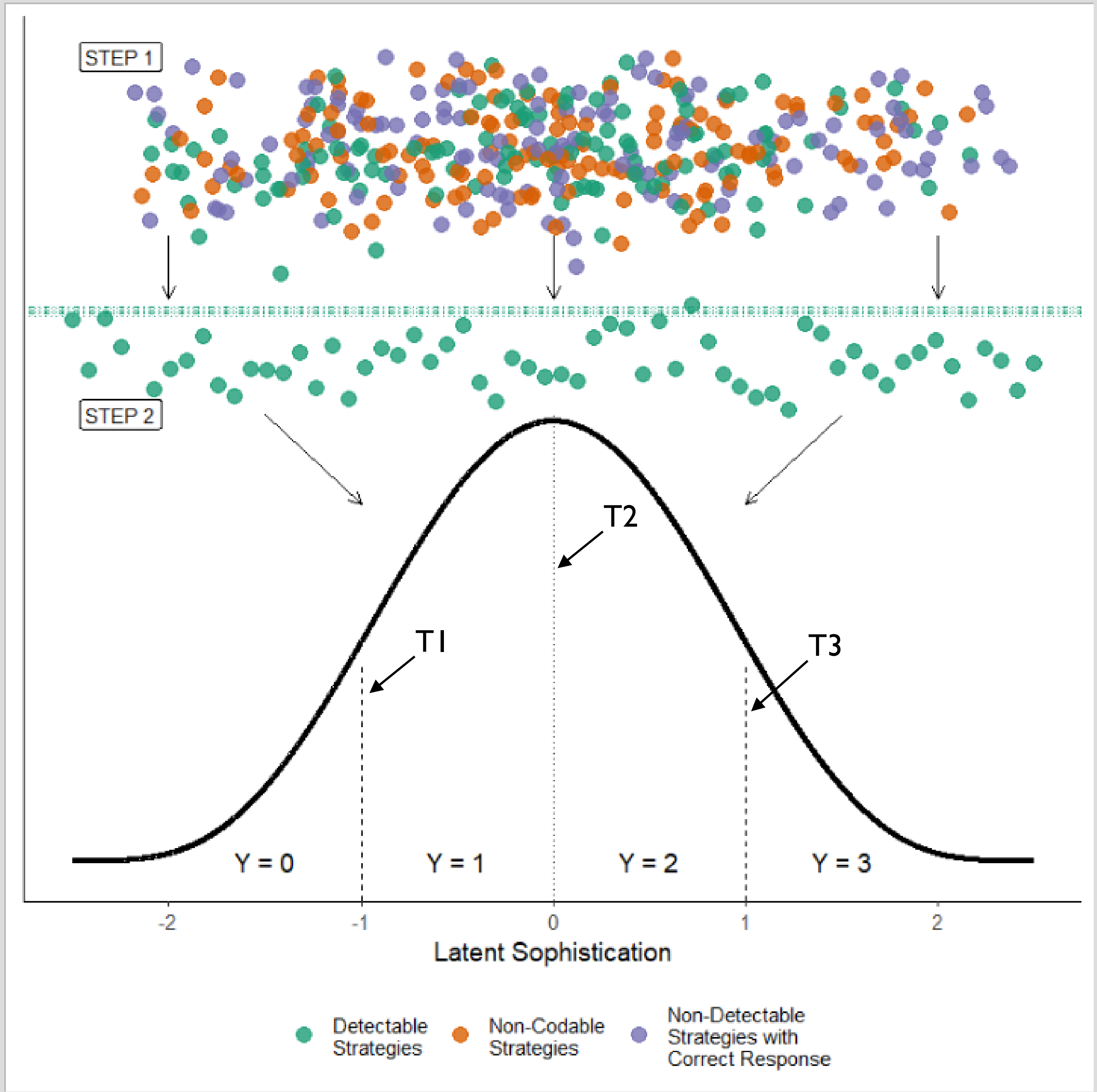
Step 1: Collect and Code Data

1) Detectable Strategies
2) Non-Codable Strategies
3) Non-Detectable Strategies with Correct Responses

Step 2 (a): Filter *Detectable Strategies* and determine sophistication level

- 0 = mean level of latent sophistication in the population of kindergarten-aged children.

Step 2 (b): Determine probability of *Non-Codable* and *Non-Detectable Strategies* based on what's left over



STATISTICAL ANALYSIS: HURDLE MODEL

Cumulative Logit Hurdle Model

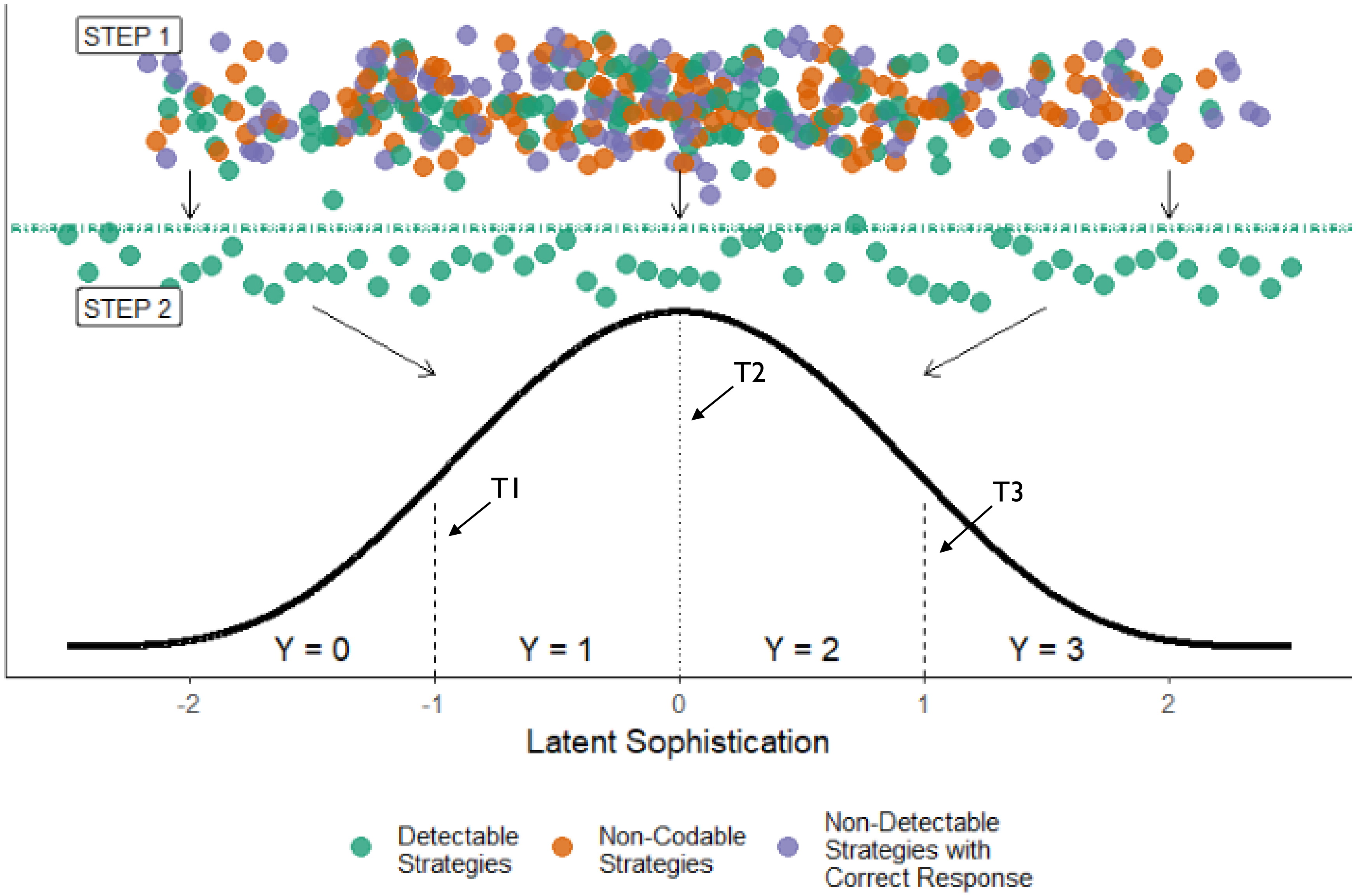
Y_{ijk} denotes one of 4 detectable strategies (0, 1, 2, 3) used by *ith* child, on the *jth* item, in *kth* classroom

π_{ijk} denotes the probability of recording a detectable strategy

We assume that non-codable and non-detectable strategies are not observable under the cumulative model:

$$CL(Y_{ijk} = H|x) = 0$$

$$Pr(Y_{ijk} = c|x_{ijk}) = \begin{cases} 1 - \pi_{ijk}, & c = H \\ \pi_{ijk} \times \underbrace{[Pr(Y_{ijk} \leq c + 1|x_{ijk}) - Pr(Y_{ijk} \leq c|x_{ijk})]}_{\text{Cumulative logit model}}, & c \neq H \end{cases}$$



$$Pr(Y_{ijk} = c|x_{ijk}) = \begin{cases} 1 - \pi_{ijk}, & c = H \\ \pi_{ijk} \times \underbrace{[Pr(Y_{ijk} \leq c + 1|x_{ijk}) - Pr(Y_{ijk} \leq c|x_{ijk})]}_{\text{Cumulative logit model}}, & c \neq H \end{cases}$$

STATISTICAL ANALYSIS: HURDLE MODEL

Cumulative Logit Hurdle Model

$$\begin{cases} 1-\left(\pi_{ijk}\middle|x_{ijk}\right), & Y_{ijk}=H \\ \left(\pi_{ijk}\middle|x_{ijk}\right)\times CL\left(Y_{ijk}\le c\middle|x_{ijk}\right), & Y_{ijk}\ne H \end{cases}$$

Probability of Detection

$$logit\left(\pi_{ijk}\right)=\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right)=\alpha+x'_{ijk}\boldsymbol{\beta}^{(d)}+u_i^{(d)}+v_j^{(d)}+w_k^{(d)}$$

$$u_i^{(d)}\sim N\left(0,\sigma_u^{(d)}\right)\quad v_j^{(d)}\sim N\left(0,\sigma_v^{(d)}\right)\quad w_k^{(d)}\sim N\left(0,\sigma_w^{(d)}\right)$$

$$\left\{\sigma_u^{(d)},\sigma_v^{(d)},\sigma_w^{(d)}\right\}\sim HalfN(0,1.5)$$

$$\alpha\sim N(0,2.0)$$

$$\boldsymbol{\beta}^{(d)}=\left\{\beta_1^{(d)}\ldots\beta_5^{(d)}\right\}\sim N(0,2.0)$$

Sophistication Given Strategy Detected

$$logit\left(\Pr(Y_{ijk}\le c)\right)=\log\left(\frac{\Pr(Y_{ijk}\le c)}{\Pr(Y_{ijk}>c)}\right)=\theta_c-\left(x'_{ijk}\boldsymbol{\beta}+u_i+v_j+w_k\right)$$

$$u_i\sim N(0,\sigma_u)\quad v_j\sim N(0,\sigma_v)\quad w_k\sim N(0,\sigma_w)$$

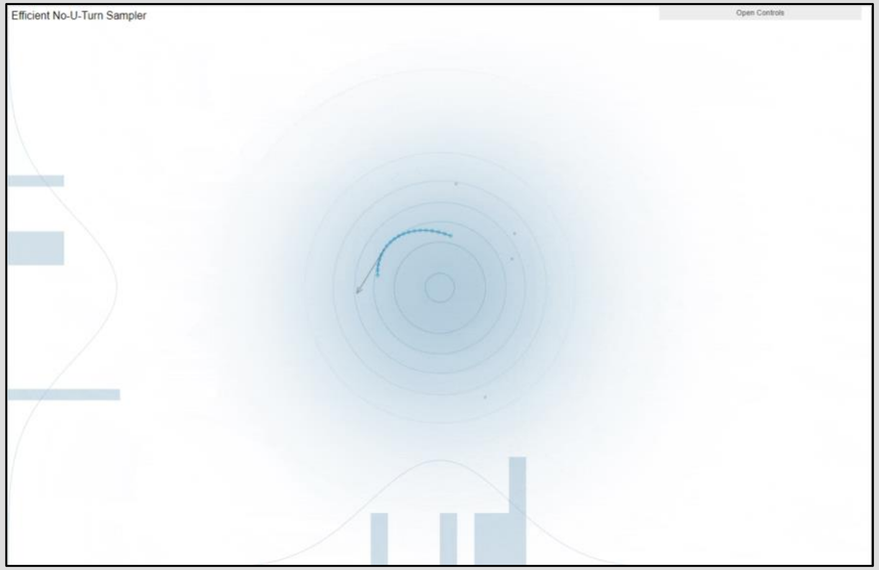
$$\{\sigma_u,\sigma_v,\sigma_w\}\sim HalfN(0,1.5)$$

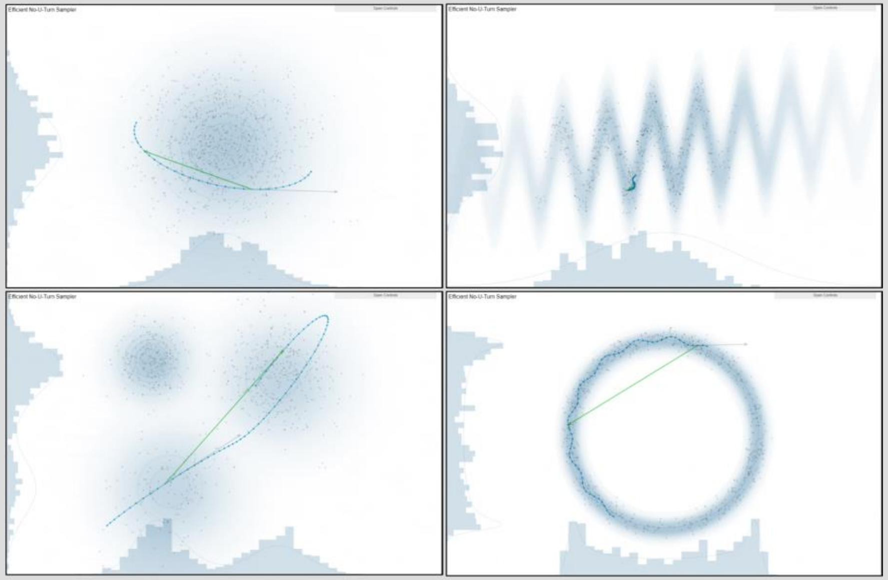
$$\theta_c\sim N(0,2.5)$$

$$\boldsymbol{\beta}=\{\beta_1\ldots\beta_5\}\sim N(0,1.5)$$

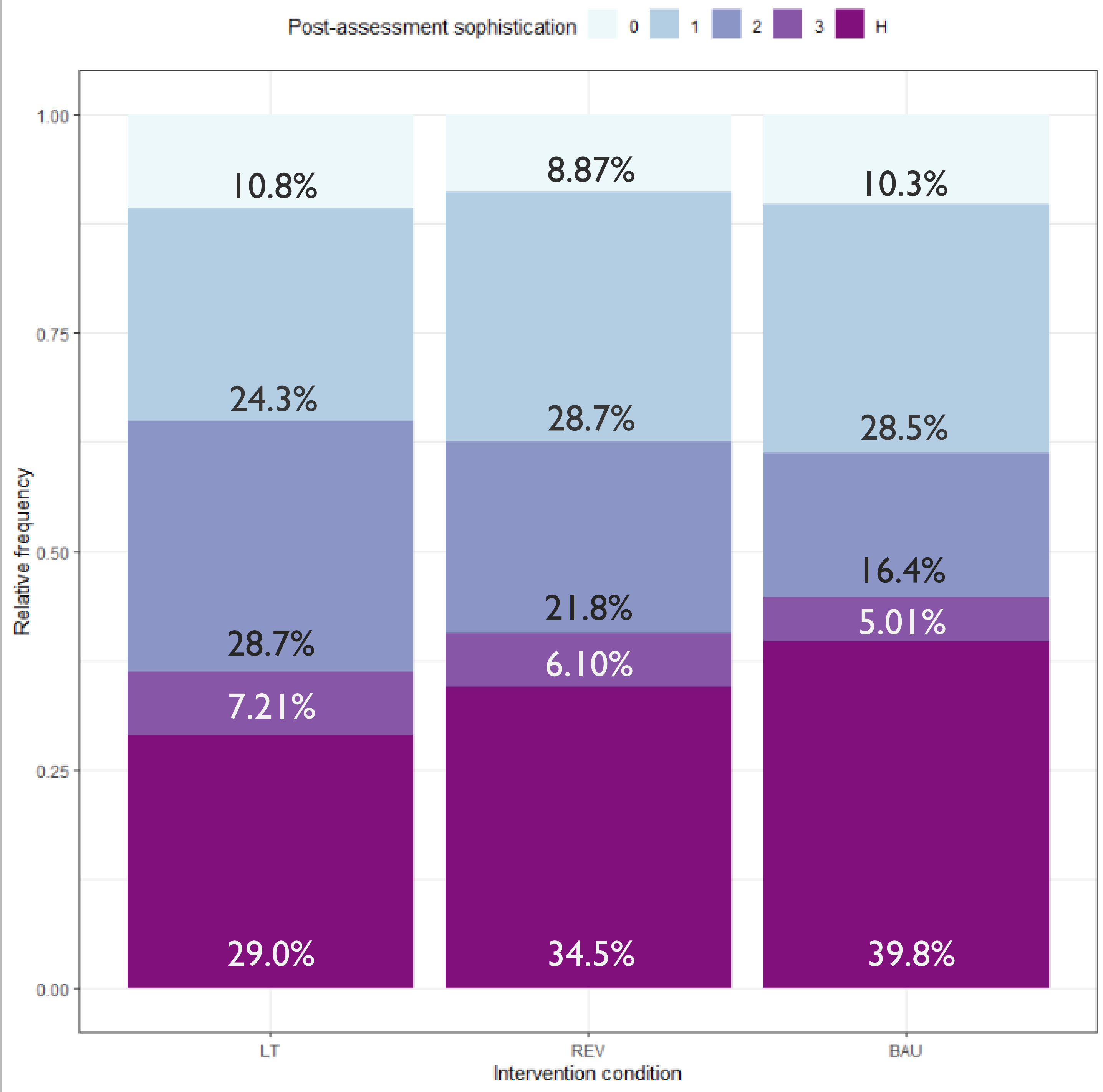
STATISTICAL ANALYSIS: ESTIMATION

- Samples were drawn using No U-Turn Sampler (NUTS) [Hamiltonian Monte Carlo Method (HMC)]
  - Automatically selects an appropriate number of leapfrog steps in each iteration in order to allow the proposals to traverse the posterior
  - Maximize the expected squared jump distance at each step to avoid random-walk behavior





DATA SUMMARY



Post-assessment sophistication across Intervention Conditions

0 - 3 = lowest to highest post-assessment sophistication strategy

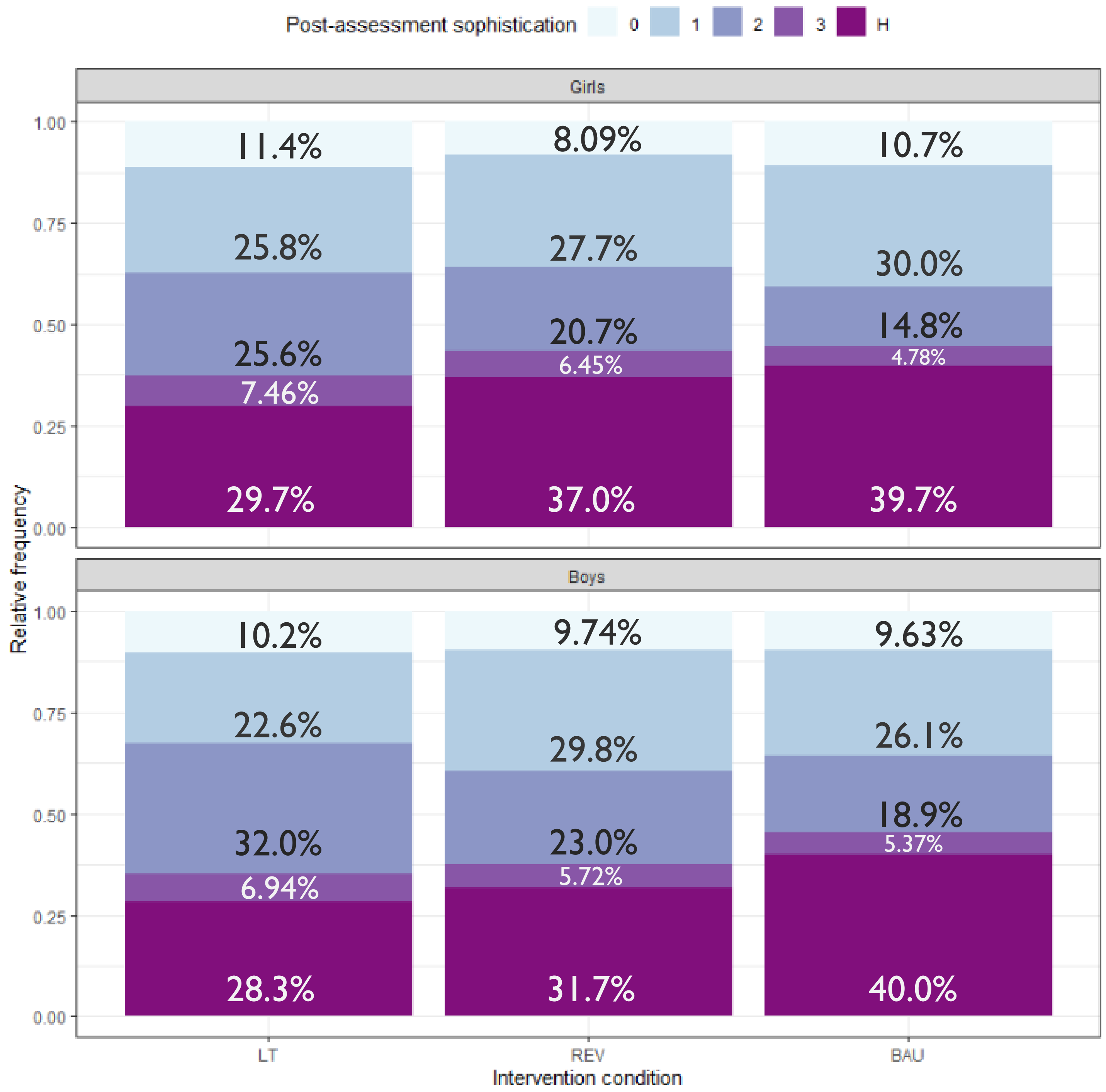
H = non-detectable (non-codable) sophistication strategy

Without the Hurdle Model?

- Losing an average of 34.43% of information within each condition

Takeaways:

- LT has the smallest proportion of non-codable & non-detectable strategies
- BAU has the largest proportion of non-codable & non-detectable strategies
- LT also has a larger proportion of more sophisticated strategies when compared to the other two conditions



Post-assessment sophistication across <u>Boys & Girls</u>

0 - 3 = lowest to highest post-assessment sophistication strategy

H = non-detectable (non-codable) sophistication strategy

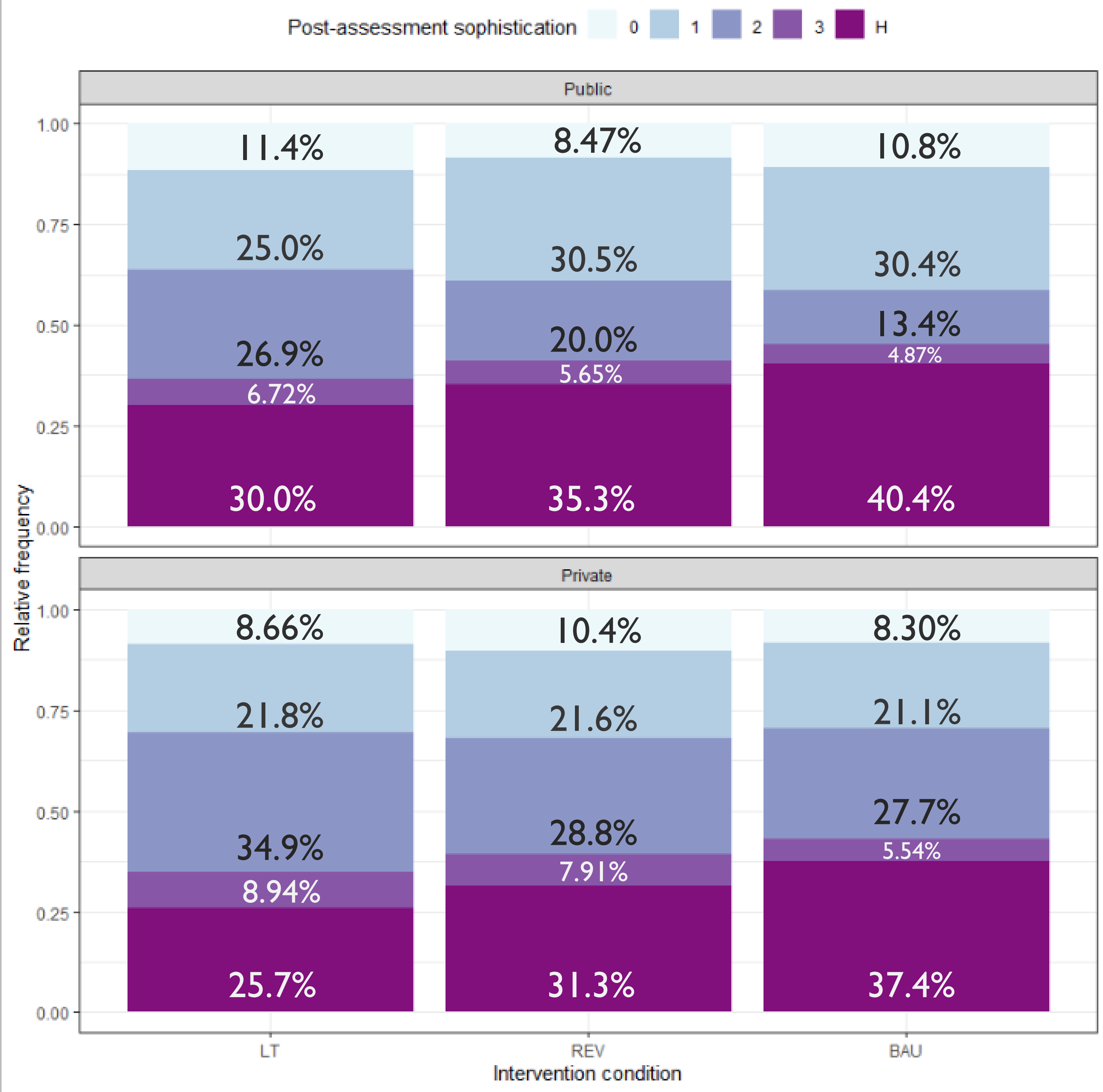
Without the Hurdle Model?

- Losing an average of 35.47% of information within *girls* within each condition
- Losing an average of 33.33% of information within *boys* within each condition

Takeaways:

- *Boys & girls* within *LT* have the lowest proportion of non-codable & non-detectable strategies

- LT has a larger proportion of more sophisticated strategies for both *boys & girls*



Post-assessment sophistication across School Type

0 - 3 = lowest to highest post-assessment sophistication strategy

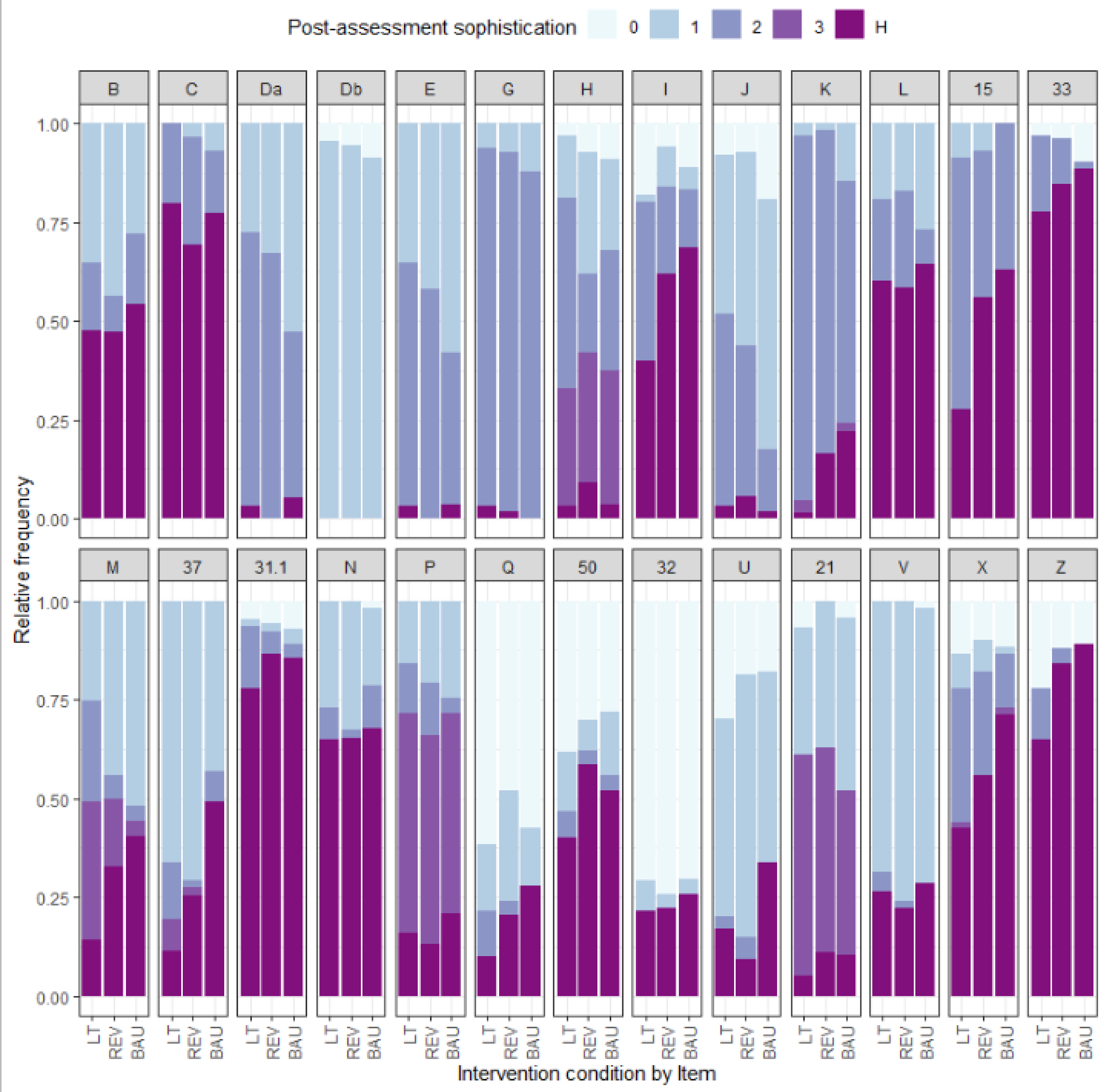
H = non-detectable (non-codable) sophistication strategy

Without the Hurdle Model?

- Losing an average of 35.23% of information within *public schools* within each condition
- Losing an average of 31.4% of information within *private schools* within each condition

Takeaways:

- Same story, different setting: LT has a larger proportion of more sophisticated strategies when compared to the other conditions with public & private schools



Post-assessment sophistication across Intervention Conditions by Item

0 - 3 = lowest to highest post-assessment sophistication strategy

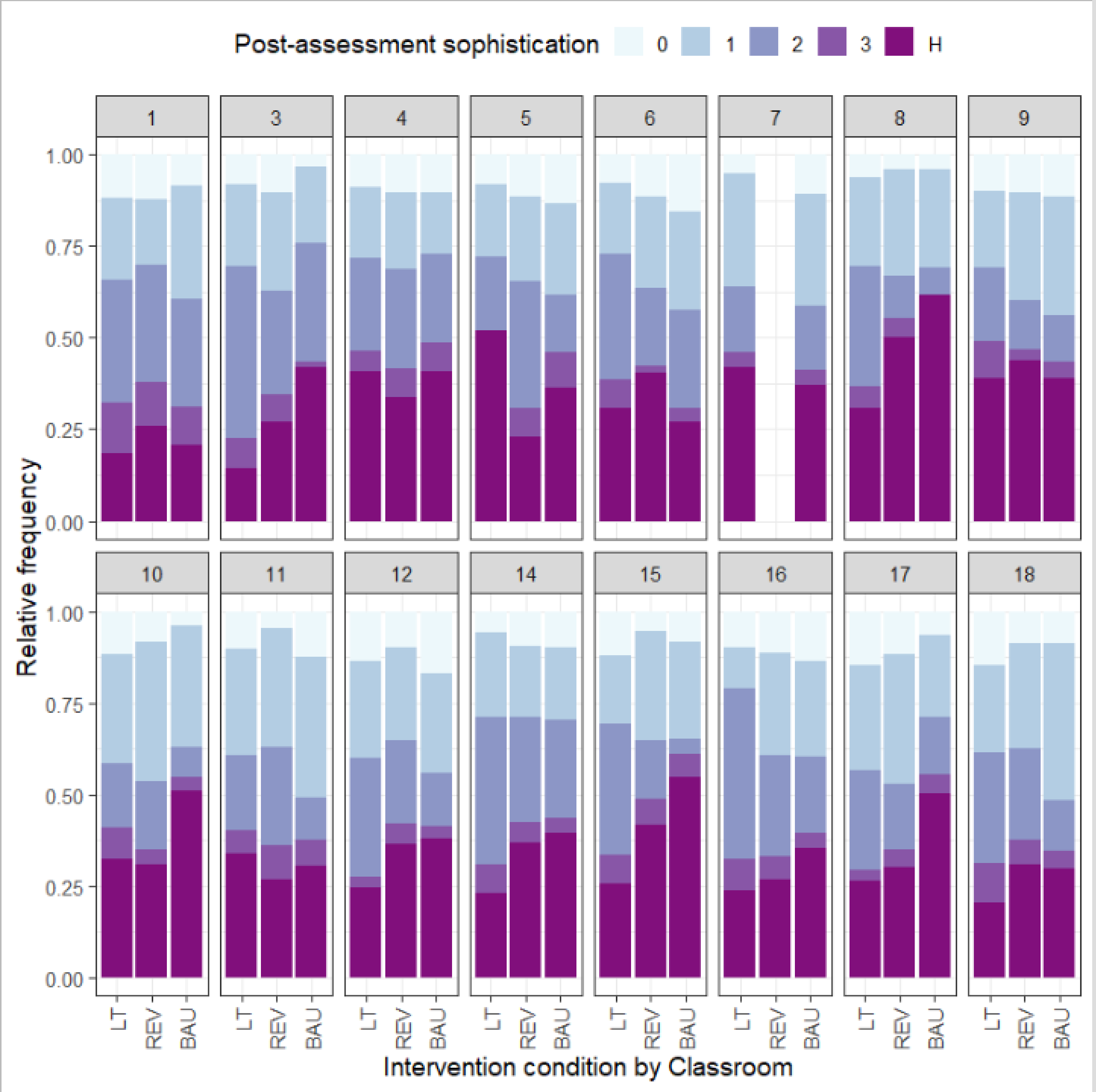
H = non-detectable (non-codable) sophistication strategy

Without the Hurdle Model?

- Massive loss of information within most items
    - ***Item 33:***
        - 83.63% information loss
    - ***Item 31.3:***
        - 83.26% information loss
    - **Item Z:**
        - 79.3% information loss

Takeaways:

- Large variability of sophistication strategies within each item regardless of codability



Post-assessment sophistication across Intervention Conditions by Classroom

0 - 3 = lowest to highest post-assessment sophistication strategy

H = non-detectable (non-codable) sophistication strategy

Without the Hurdle Model?

- 34.32% mean information loss across all classes

Takeaways:

- Students in the LT condition within the varying classes follow previous trends when compared to the other strategies:
  - ***LT produces more codable strategies***
  - ***LT produces more sophisticated strategies***

RESULTS

MODEL SELECTION

Model* (fixed effects)	WAIC	LOOIC
1) Pre-Soph	8969.35	8970.36
2) Pre-soph + Intervention	8943.37	8944.16
3) Pre-soph x Intervention	8945.03	8945.84
4) Pre-soph + Intervention + Sex	8943.96	8944.74
5) Pre-soph x Intervention + Sex	8947.53	8948.32
6) Pre-soph + Intervention + Sex + Private	8941.59	8942.36
7) Pre-soph x Intervention + Sex + Private	8943.37	8944.18
8) Pre-soph + Intervention + Sex + Private + Private x Intervention	8944.4	8945.23

*Random Effects: Child ID + Item + Class

- The *lowest* _IC indicates the "best" model
  - ***Widely Applicable Information Criteria*** (WAIC): mean log likelihood function over the posterior distribution + correction

$$WAIC = -2\sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{j=1}^{S}p(x_i|\theta_j)\right) + 2\sum_{i=1}^{n}\left(V_{j=1}^{S}\log p(x_i|\theta_j)\right)$$

  - ***Leave-One-Out Information Criteria*** (LOOIC): mean fit of the model per data point over posterior draws

$$LOOIC = \sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{j=1}^{S}p(x_i|x_{-i},\theta_j)\right)$$

n = sample size, S = number of draws

DETECTING CODABLE STRATEGIES

--	Mean	SD	2.5%	97.5%
Pre-sophistication	0.25	0.11	0.04	0.46
LT (w.r.t BAU)	0.79	0.16	0.48	1.09
REV (w.r.t BAU)	0.40	0.16	0.09	0.73
Sex (M = 1, F = 0)	0.12	0.14	-0.15	0.39
School Type (Private = 1, Public = 0)	0.17	0.28	-0.37	0.75

Practical Conclusions:

- For each SD increase in pre-sophistication Rasch score, the odds of detecting a strategy increase by ~1.3x (exp(0.25))

- LT group had ~2.2x (exp(0.79)) greater odds of detecting a strategy when compared to BAU groups

- LT children are more likely to use a conceptually meaningful and relevant strategy at post-assessment relative to BAU

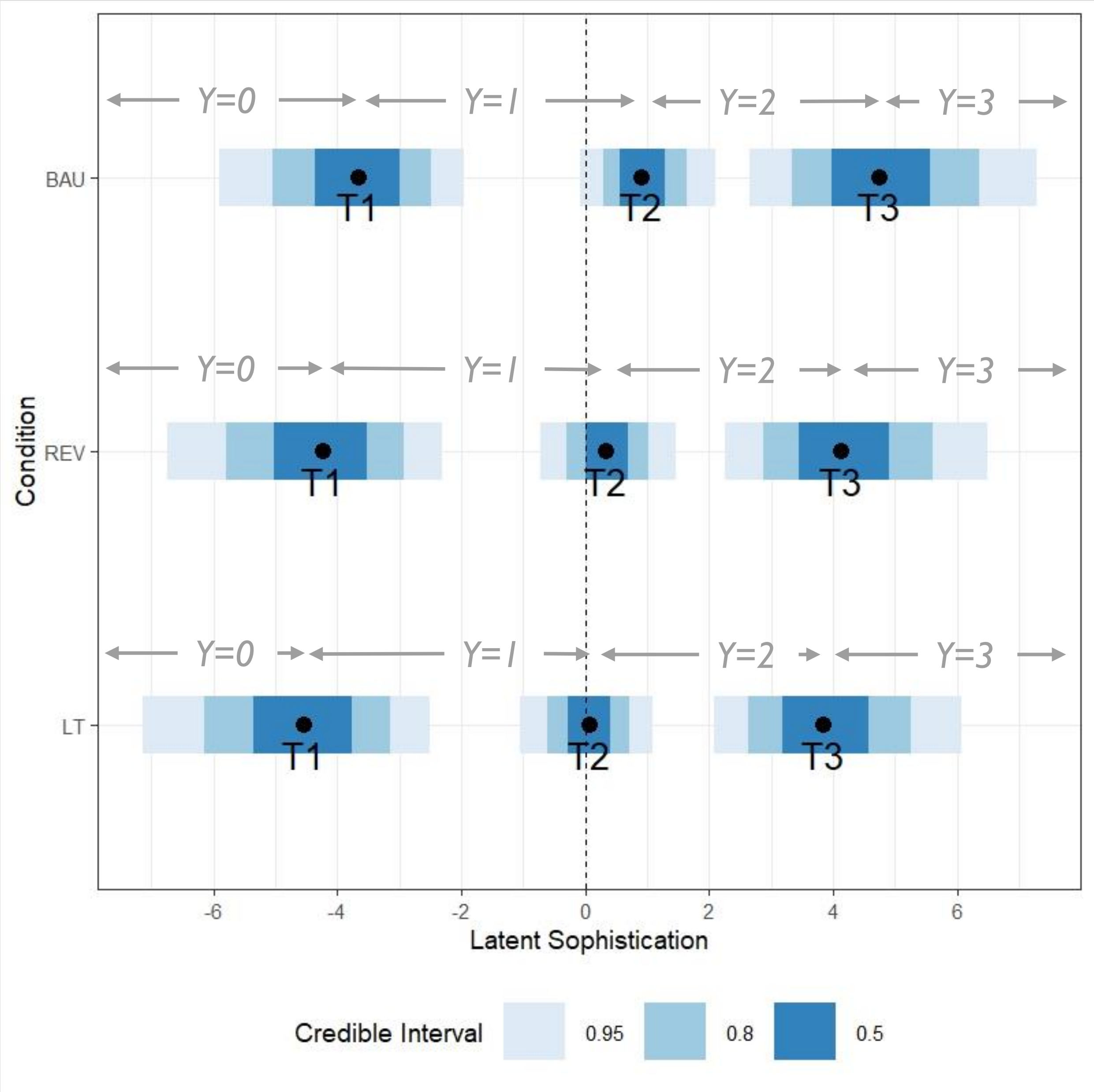
FIXED EFFECTS WHEN A STRATEGY IS DETECTED

--	Mean	SD	2.5%	97.5%
Pre-sophistication	0.38	0.13	0.17	0.67
LT (w.r.t BAU)	0.89	0.25	0.46	1.42
REV (w.r.t BAU)	0.59	0.20	0.25	1.04
Sex (M = 1, F = 0)	0.04	0.13	-0.21	0.30
School Type (Private = 1, Public = 0)	0.51	0.21	0.15	0.99

Practical Conclusions:

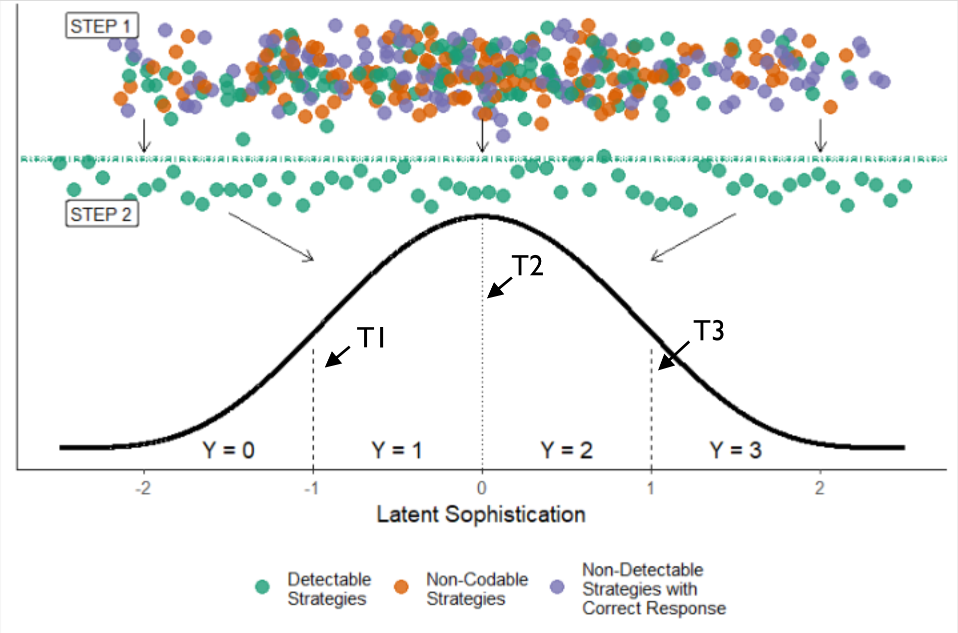
- Students in the *LT & REV* group have significantly higher odds of utilizing more sophisticated strategies than *BAU*
- *Boys* and *girls* do not differ significantly in the sophistication of their strategies when a strategy is detected.
- Students in *private schools* have significantly higher odds of utilizing more sophisticated strategies than *public schools*

SOPHISTICATION THRESHOLDS WHEN A STRATEGY IS DETECTED



Practical Conclusions:

- Average LT students were more likely to deploy a more sophisticated strategy relative to the average BAU or REV students.
- For example, the average LT student was making the transition from Y=1 to Y=2 (T2). Whereas, the average student in the BAU condition was likely still in Y=1.



Model* (fixed effects)	P(LT > REV | Detection)	P(Higher Chance of Detectable Strategy in LT)
1) Pre-Soph	--	--
2) Pre-soph + Intervention	97.73	99.28
3) Pre-soph x Intervention	97.80	99.29
4) Pre-soph + Intervention + Sex	97.57	99.15
5) Pre-soph x Intervention + Sex	97.87	99.32
6) Pre-soph + Intervention +Sex + Private	97.85	99.33
7) Pre-soph x Intervention + Sex + Private	98.23	99.5
8) Pre-soph + Intervention +Sex + Private + Private x Intervention	96.45	98.47

*Random Effects: Child ID + Item + Class
Hyper-parameters: chains = 3, warmup = 1000, iter = 5000

Practical Conclusions:

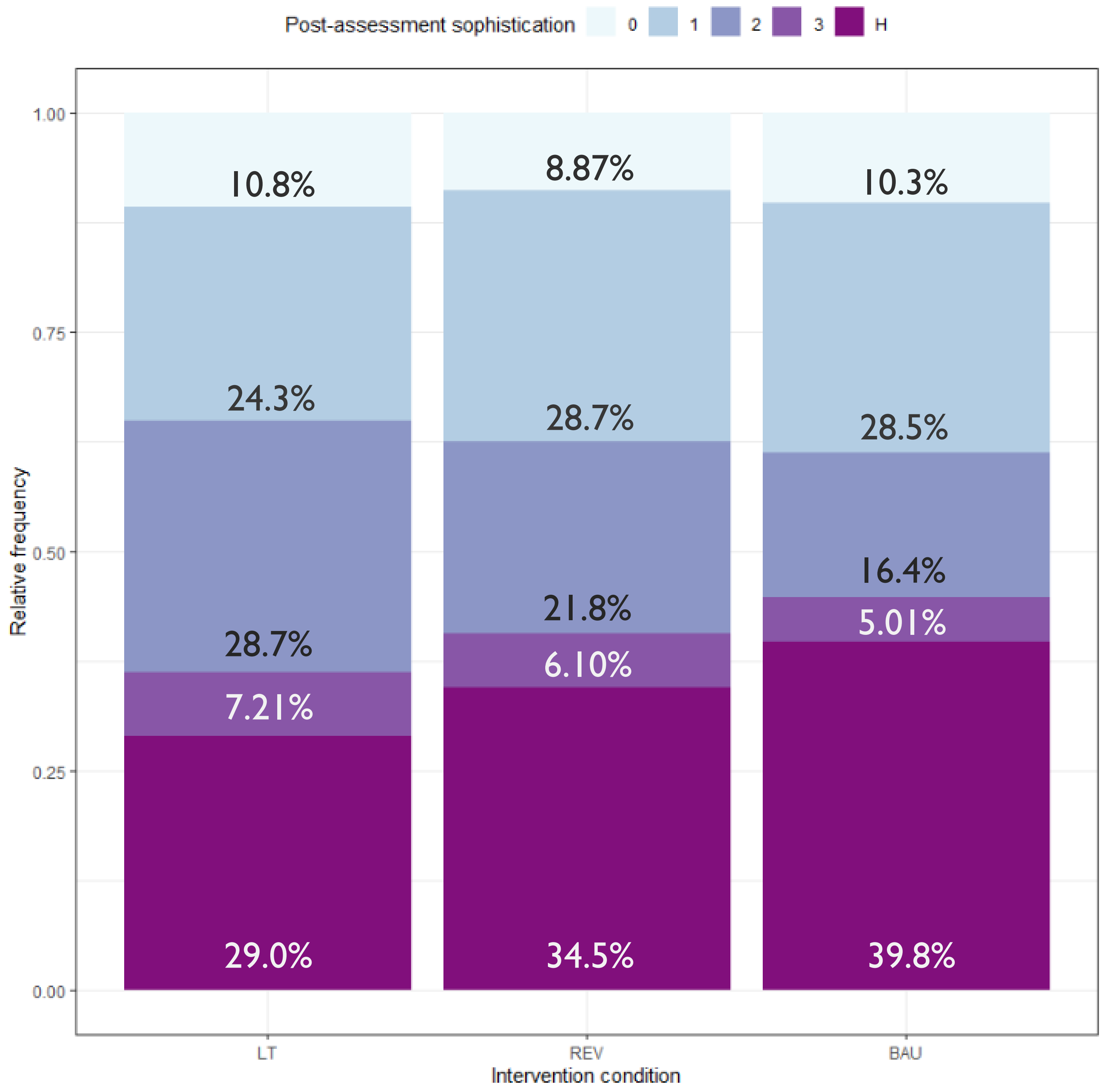
- Given a strategy was detected, there is 97.85% posterior probability that the LT students deploys a more sophisticated strategy relative to REV peers.
- There is 99.33% posterior probability that LT students will have a higher chance of using detectable strategies relative to REV peers.
- **Take Home:** Regardless of the model specified, the posterior probability that LT students use more sophisticated strategy relative to their REV peers is above 96%.

RESULTS SUMMARY

RESULTS SUMMARY

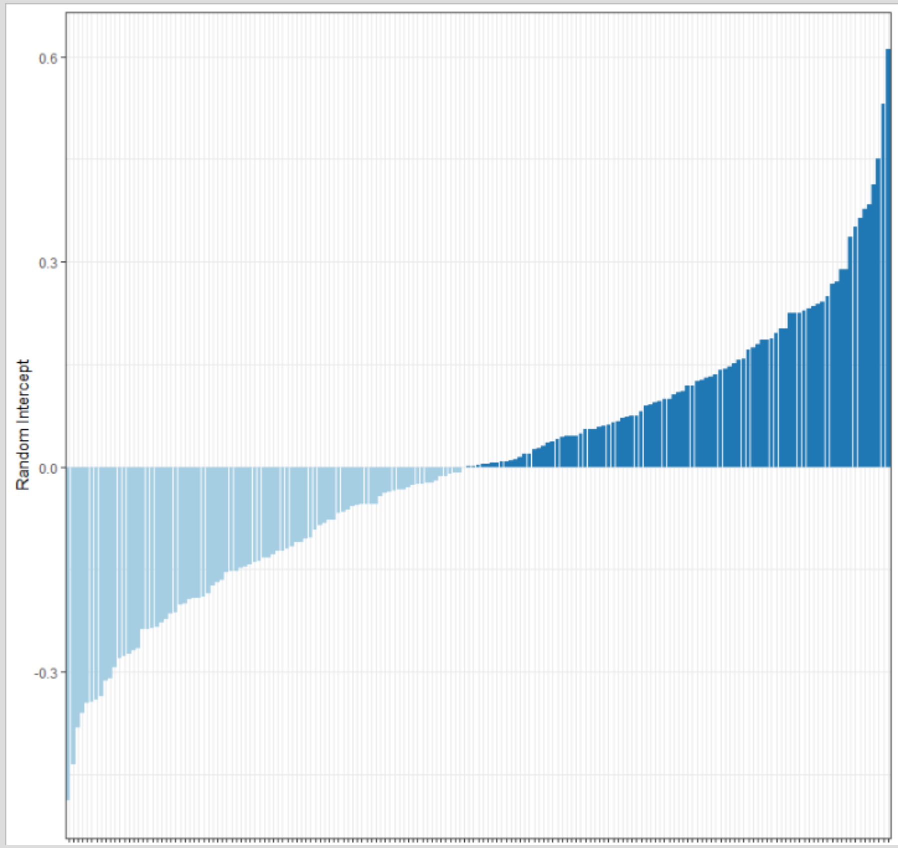
Research Question 1: Given a strategy was detected, do children in the Learning Trajectories group use more sophisticated strategies relative to their peers in two counterfactual conditions?

A: *Yes*. Given a strategy was detected, there is 97.85% probability that the LT students used a more sophisticated strategy relative to their REV peers.



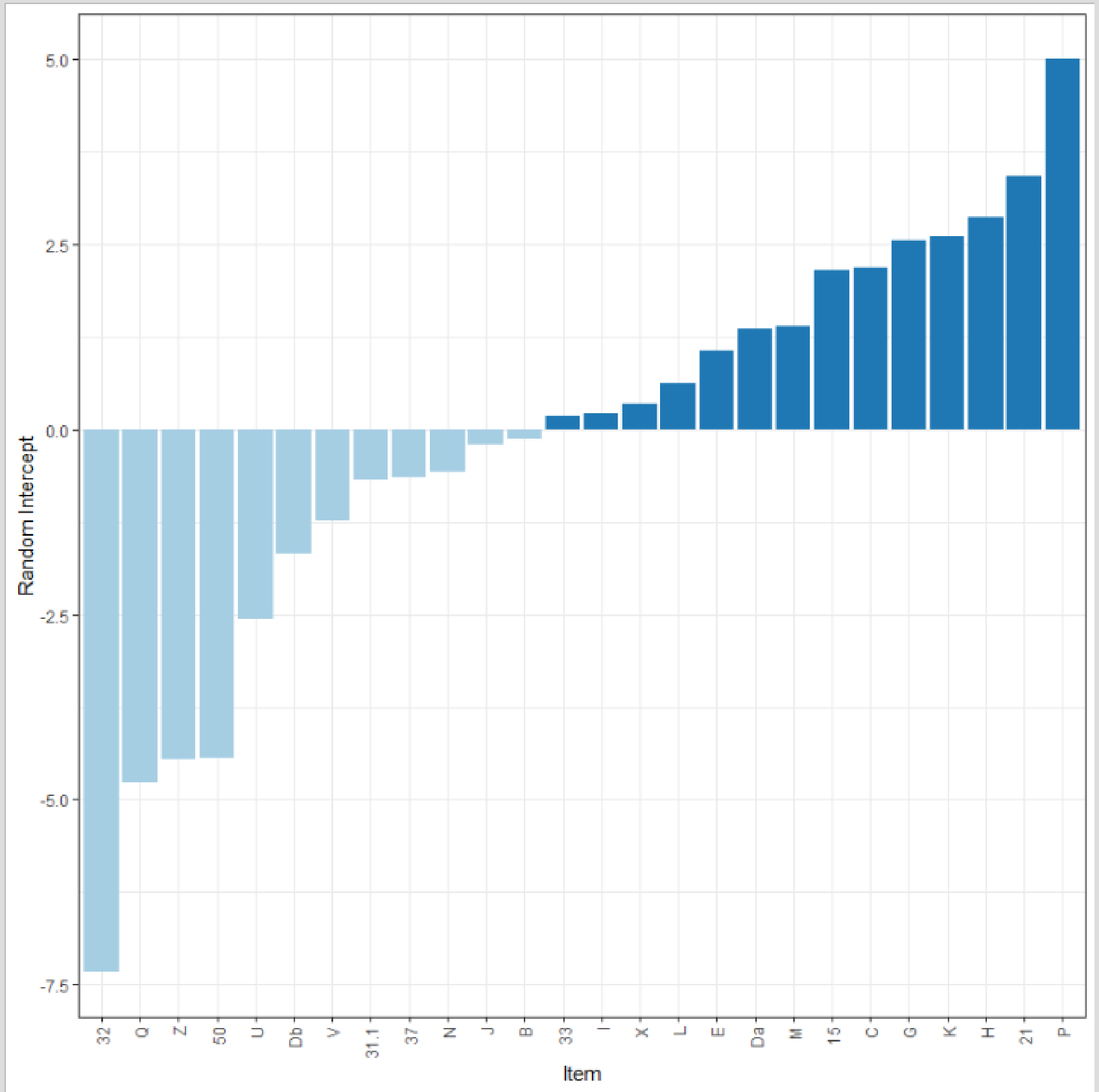
RANDOM INTERCEPTS

- *Random intercepts* measure how different *items*, *classrooms*, and *students* deviate from the population mean latent sophistication
  - **Negative R.I.** → Lower sophistication than population mean
  - **Positive R.I.** → Higher sophistication than population mean
  - **~ 0 R.I.** → ~ Around population mean sophistication
    - "Typical" population *item*, *classroom*, or *student*
      - *Random effects = 0*
    - Interpretations of fixed effects (e.g., *intervention effects*) are conditional on all random effects being 0
      - **Typical** student from a typical classroom evaluating a typical item



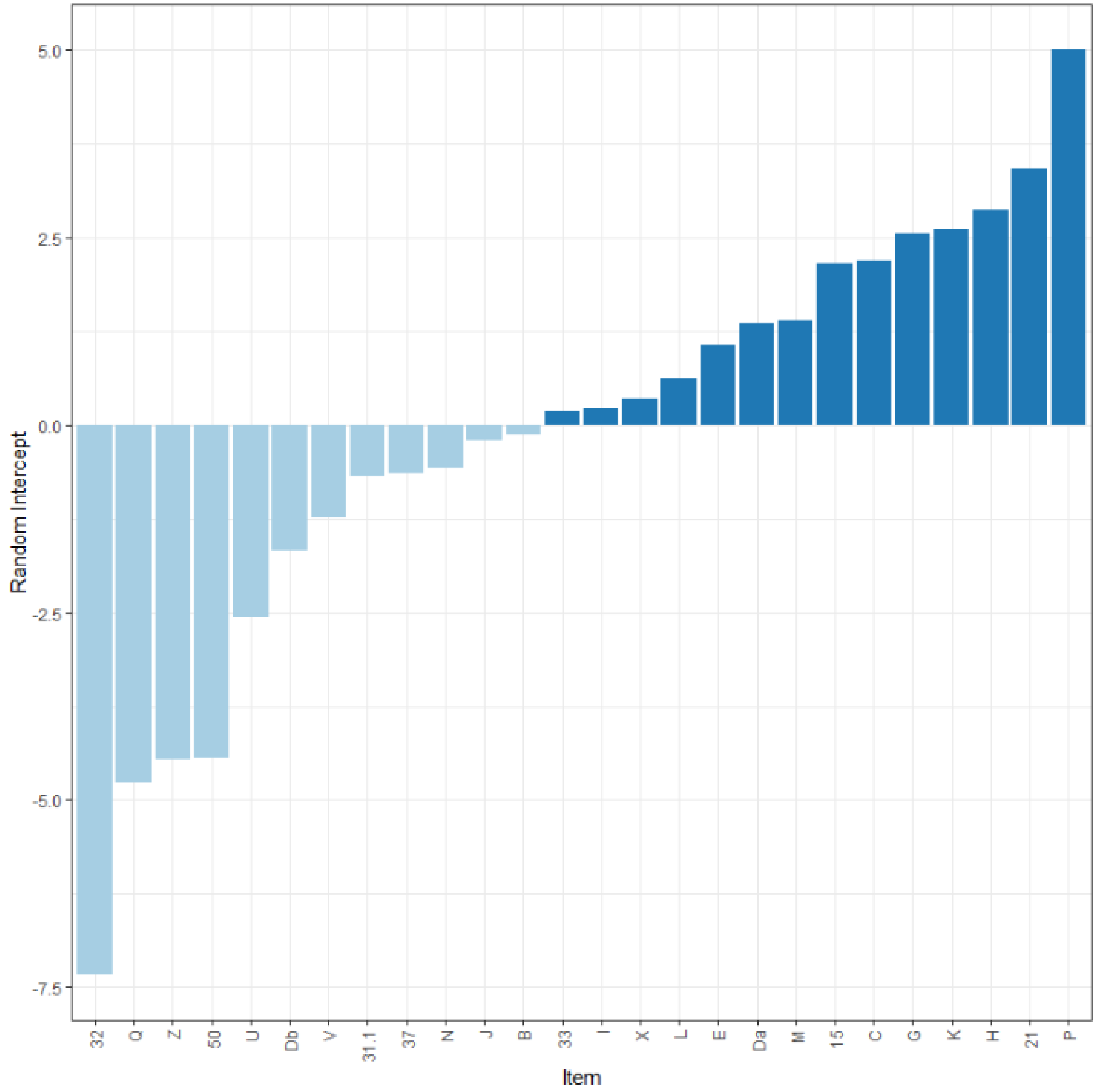
RESEARCH QUESTIONS

Research Question 2a: What can ***item*** random effects tell us about strategy preference?



Item	Random Intercept	Item	Random Intercept
32	-7.32	*33*	0.19
Q	-4.76	*I*	0.23
Z	-4.46	*X*	0.35
50	-4.43	*L*	0.62
U	-2.56	*E*	1.06
Db	-1.68	*Da*	1.37
V	-1.23	*M*	1.39
31.1	-0.68	*15*	2.15
37	-0.64	*C*	2.19
N	-0.58	*G*	2.55
J	-0.20	*K*	2.62
B	-0.12	*H*	2.87
-	-	*21*	3.42
-	-	*P*	5.00

- Highlighted items are within 0.25 of the mean (0)
  - "Typical" items in terms of length measurement strategy sophistication

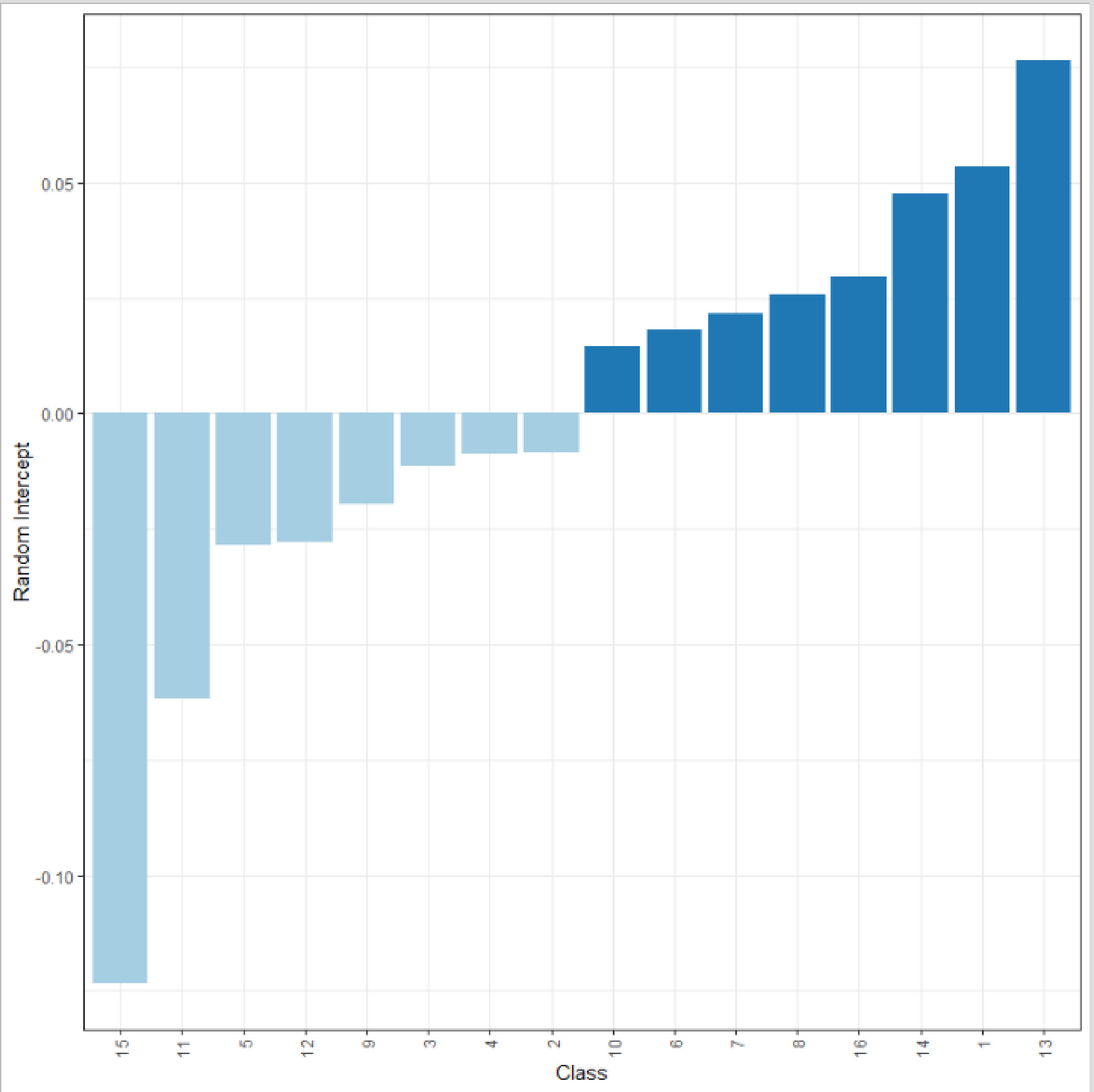


Practical Conclusions:

- “Typical” items cover LQR, LDC, EE, & LURR thinking
- High sophistication items address EE thinking
- Low sophistication items address LURR+ thinking
- Notably, there is large deviation between items

RESEARCH QUESTIONS

Research Question 2b: What can ***classroom*** random effects tell us about strategy preference?



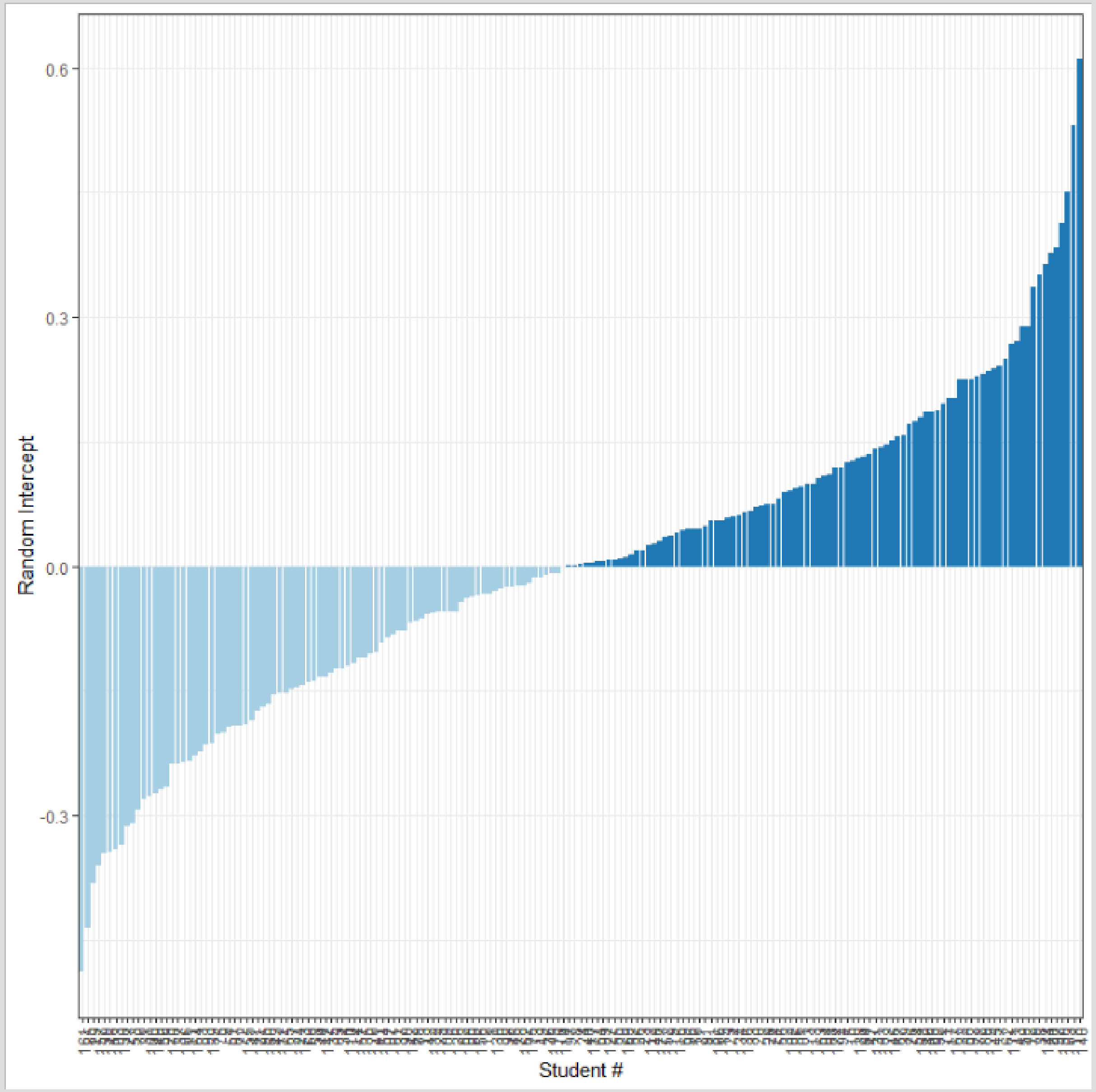
Class	Random Intercept	Class	Random Intercept
15	-0.12	*10*	0.01
11	-0.06	*6*	0.02
5	-0.03	*7*	0.02
12	-0.03	*8*	0.03
9	-0.02	*16*	0.03
3	-0.01	*14*	0.05
4	-0.01	*1*	0.05
2	-0.01	*13*	0.08

Practical Conclusions:

- There is very little deviation in latent sophistication between classes
- Classrooms likely don't play a major roll in latent sophistication

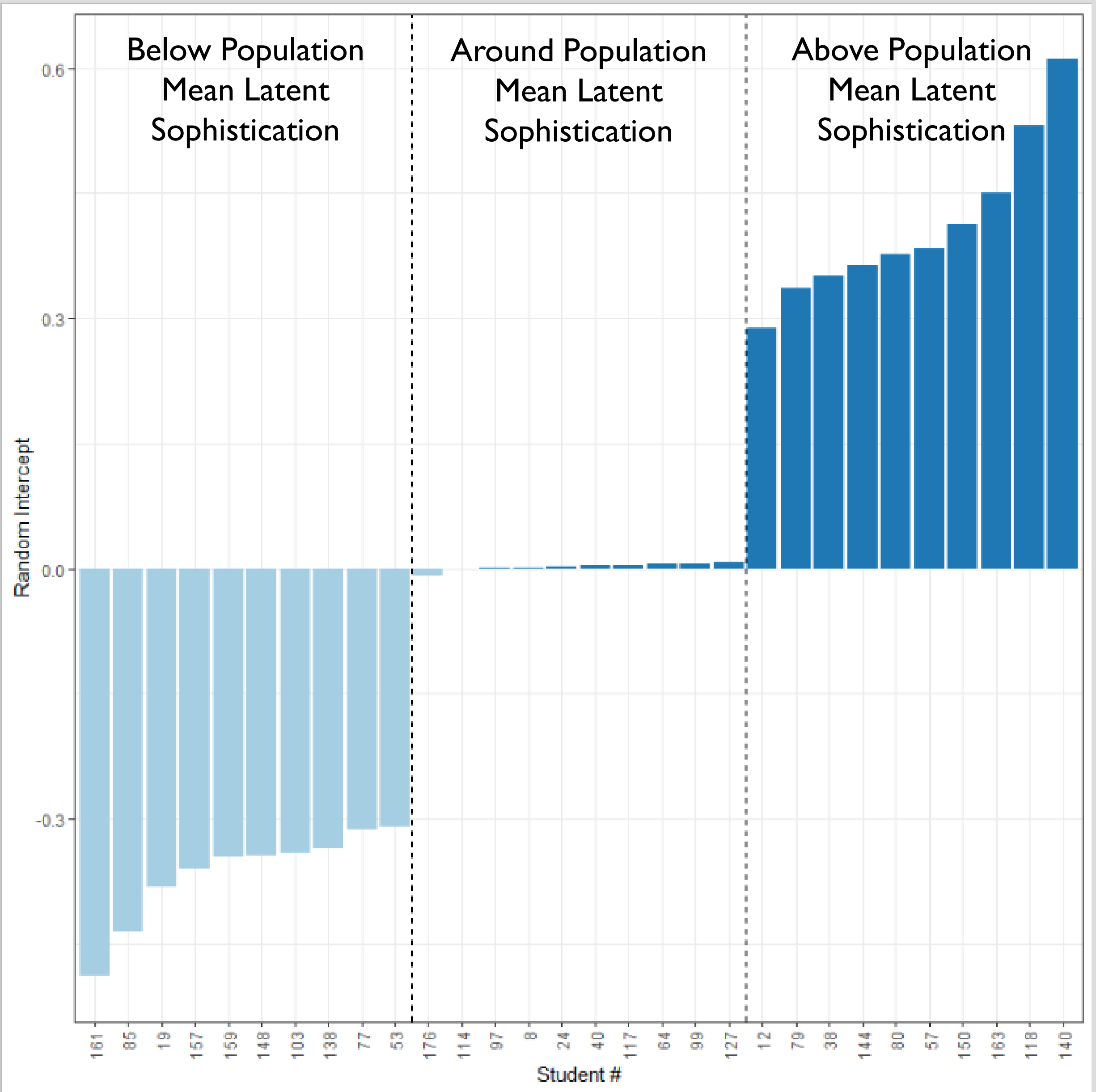
RESEARCH QUESTIONS

Research Question 2b: What can ***student*** random effects tell us about strategy preference?



RESEARCH QUESTIONS

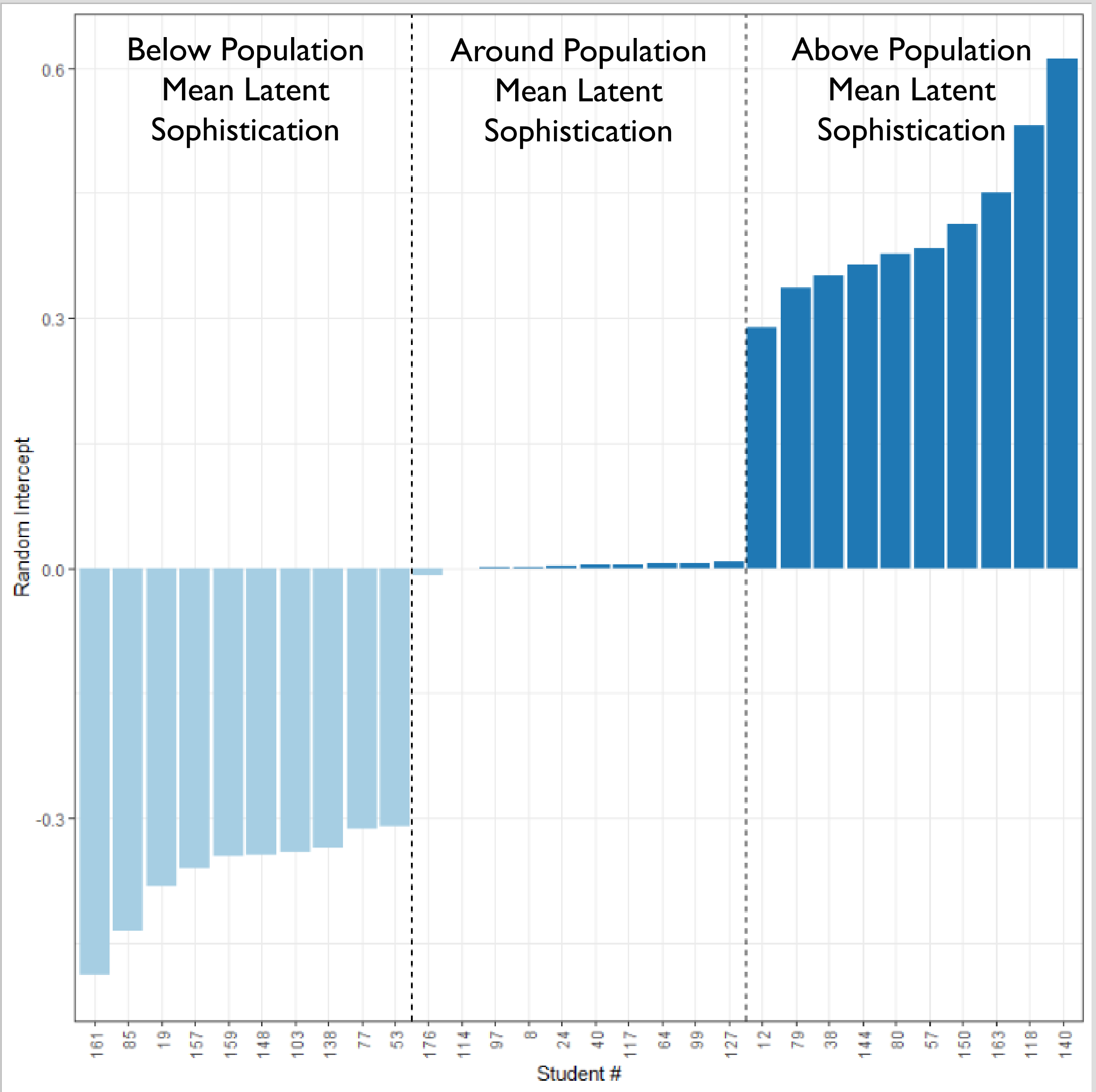
Research Question 2c: What can ***student*** random effects tell us about strategy preference?



Below Population Mean Latent Sophistication		Around Population Mean Latent Sophistication		Above Population Mean Latent Sophistication	
Student #	**R.I.**	**Student #**	**R.I.**	**Student #**	**R.I.**
161	-0.49	176	-0.01	12	0.29
85	-0.44	114	0.00	79	0.34
19	-0.38	97	0.00	38	0.35
157	-0.36	8	0.00	144	0.36
159	-0.34	24	0.00	80	0.38
148	-0.34	40	0.00	57	0.38
103	-0.34	117	0.00	150	0.41
138	-0.34	64	0.01	163	0.45
77	-0.31	99	0.01	118	0.53
53	-0.31	127	0.01	140	0.61
Bottom 10 Students		Middle 10 Students “Typical Students”		Top 10 Students	

RESEARCH QUESTIONS

Research Question 2c: What can ***student*** random effects tell us about strategy preference?



Practical Conclusions:

- It *is* possible to score every student in terms of their sophistication and use these scores like Rasch scores in correctness-based assessments.

RESEARCH QUESTIONS

Research Question 3: How do we best model strategy preference, given a large portion of strategies do not fall onto the existing research-based sophistication scale?

A1: When a strategy can be detected, Pre-sophistication level, Experimental Condition (LT, REV, BAU), & School Type (Public, Private) contribute the most to sophistication.

A2: Regardless of the model specified, the posterior probability that children in the LT condition use more sophisticated strategy vs. REV condition is above 96%.

A3: There is a methodological advantage to using the hurdle model due to the differential probabilities of a detectable strategy by Experimental Condition, Pre-Sophistication, and School Type (less information loss, more power)

QUESTIONS?